

A Markov Chain based Ensemble Method for Crowdsourced Clustering

Sujoy Chatterjee¹, Enakshi Kundu¹ and Anirban Mukhopadhyay¹

¹Department of Computer Science and Engineering, University of Kalyani, Nadia – 741235, India

E-mail: {sujoy, anirban}@klyuniv.ac.in, enakshikundu@gmail.com

Abstract

In presence of multiple clustering solutions for the same dataset, a clustering ensemble approach aims to yield a single clustering of the dataset by achieving a consensus among the input clustering solutions. The goal of this consensus is to improve the quality of clustering. It has been seen that there are some image clustering tasks that cannot be easily solved by computer. But if these images can be outsourced to the general people (crowd workers) to group them based on some similar features, and opinions are collected from them, then this task can be managed in an efficient manner and time effective way. In this work, the power of crowd has been used to annotate the images so that multiple clustering solutions can be obtained from them and thereafter a Markov chain based ensemble method is introduced to make a consensus of multiple clustering solutions.

Introduction

Clustering is a common unsupervised learning method, which is used to find hidden patterns or groupings in data. These groups are termed as clusters, and there are different types of clustering techniques (Berkhin 2002; Jain, Murthy, and Flynn 1999) that partition the dataset in different ways. In unsupervised classification, known as clustering, it is not known beforehand how the data is grouped. There are some drawbacks in all existing clustering techniques. Few clustering methods are also very sensitive to the initial clustering settings. In cluster analysis, the evaluation of results is generally done using of cluster validity indexes (Mukhopadhyay, Maulik, and Bandyopadhyay. 2015; Davies and Bouldin 1979; Rand 1971; Dunn 1974; Hubert and Arbie 1985) which are employed to measure the quality of clustering results. However, there is no cluster validity index that impartially evaluates the results of any clustering algorithm. So, to combine the multiple diverse clustering solutions for achieving an improved clustering, an ensemble of clustering solutions is needed.

Although over the years numerous clustering ensemble algorithms (Strehl and Ghosh 2002; Ayad and Kamel 2008; Chatterjee and Mukhopadhyay 2013) have been proposed to

solve different issues related to cluster analysis. These algorithms have several benefits and pitfalls. Moreover, there are some hard image clustering tasks that cannot be solved by a computer in limited amount of time. But if this large image clustering task can be outsourced to numerous crowd workers (Brabham 2013; Hovy et al. 2013; Lease 2011) and clustering solutions can be obtained from them, then the task can be solved in a very effective way through a cluster ensemble approach.

In this paper, an online platform is designed to collect the clustering solutions from the crowd workers over some tricky images and a Markov chain based ensemble technique is proposed to find a robust consensus from multiple crowd based clustering solutions. This proposed scheme might be a better utilization of enormous human power that can easily solve the large image clustering task.

Problem Formulation

Let $Z = \{z_1, z_2, \dots, z_o\}$ be a set of o data objects and there are p crowd workers each of whom provides an individual clustering solution. So, $E = \{e_1, e_2, \dots, e_p\}$ denotes a set of p input clustering solutions obtained from them. Now in this problem, the number of clusters is assumed to be fixed in all the clustering solutions. So, the objects are partitioned into n clusters denoted as $C = \{c_1, c_2, \dots, c_n\}$. The objective of the problem is to find out the ensemble solution τ from these multiple input clustering solutions $E = \{e_1, e_2, \dots, e_p\}$ so that the similarity between τ and all of the input clustering solutions of E is maximized.

Proposed Model

We have made an online platform where we have posted some tricky images (that is hard for a computer to distinguish) and solicited crowd opinions to cluster those images based on their similarity. As this is a fixed size clustering problem so the number of clusters has also been posted. As various crowd workers might group the images from different viewpoints, diverse clustering solutions can be obtained. In this way, the artificial dataset is created. Fig. 1 shows the snapshot of a question for image clustering task posted to the crowd workers. Here, the question contains 7 images (a, b, c, d, e, f, g) which are asked to be clustered into 4 groups based on some similarities in features. After obtaining the clustering solutions from them, the proposed ensemble method is

* Works in Progress, Fourth AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2016), Austin, TX, USA.
Copyright © 2016, Retained by the authors. All rights reserved.

applied in order to make more robust consensus clustering solution.

Markov Chain based Ensemble Technique

Cluster labels given by the crowd workers are very symbolic, i.e., two identical clustering solutions might appear different for using different cluster labels. To resolve this label correspondence problem, here, we have chosen the standard label to be that of a reference clustering solution (a clustering solution which is most similar to the rest of the other solutions) by using Adjusted Rand Index (ARI) (Hubert and Arbie 1985) as a similarity measure. The solution assigned the maximum weight (based on ARI), is chosen as the reference partition e_r according to whose labeling all the other clustering solutions are relabeled.

The clustering ensemble problem can be solved using a Markov chain that is specified by a set of states $N = \{1, 2, \dots, n\}$ and a $n \times n$ transition matrix T , each of whose entries is a non negative real number in $[0, 1]$ representing a probability. To solve this problem, object-wise transition matrix is formed and the following steps are carried out to find the consensus solution.

- **Step 1.** In the first step, the number of possible states is found. Here, the possible states mean the unique clusters. So, if there are n clusters in the clustering solutions, the number of states will be n .
- **Step 2.** The weight of each clustering solution is measured by the average similarity of it with rest of the solutions in terms of ARI. Here normalized weight $w' = \{w'_{e_1}, w'_{e_2}, \dots, w'_{e_p}\}$ is used for p clustering solutions.
- **Step 3.** We construct a transition matrix T of order $n \times n$, whose elements T_{ij} denotes the probability of transition from state i to state j for each value i, j of data object $y \in Z$. At first the transition matrix is initialized with zeros but it is then modified as described below.

To compute the different cell values of the transition matrix (for a particular object) from a set of input clustering solutions, the label given by one clustering solution is taken initially as reference and the label provided by other solutions are considered. In this step, the normalized weight of the other clustering solutions are summed up to compute the transition probability.

Now this step is computed $\forall e_m \in E$ keeping the reference clustering solution e_k fixed. In this way, after completion of this step for a particular reference clustering solution, another clustering solution is taken as the reference and the rest of the solutions are considered for it to determine the values of the transition matrix. Note that, now as all the labels are already standardized, to form the transition matrix, label of each of the clustering solution should be compared with label of rest of the solutions.

- **Step 4.** Make the transition matrix T ergodic as follows:

$$T_{ij} = T_{ij} + \frac{1 - \sum_{j=1}^n T_{ij}}{n}, \forall i, j \in \{1, 2, \dots, n\}$$

- **Step 5.** Define a stationary distribution array sd_y of order $1 \times n$, where n denotes the number of clusters. Initially, the distribution is taken as uniform distribution.

Then $sd_y * T$ is computed for m times so that sd_y reaches a converging point. Finally the cluster label for which the distribution becomes maximum is selected as the final cluster label of the corresponding object.

In this way, the same steps are repeated for all the objects and final clustering solution is achieved.

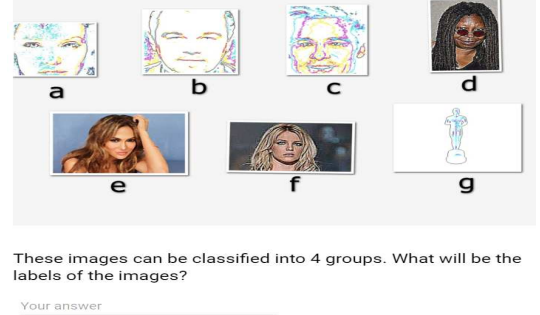


Figure 1: Snapshot of the first question posted to crowd workers.

Experimental Design and Results

We performed experiments on real-life datasets (obtained from UCI Machine Learning Repository) to find the efficacy of the proposed method and compared it with that of four well-known existing cluster ensemble algorithms, namely CSPA, HGPA, MCLA (Strehl and Ghosh 2002) and BCE (Wang, Shan, and Banerjee 2009). For image clustering task 25 crowd workers provided their solutions. The adopted performance metrics are ARI, Rand Index (RI) (Rand 1971), Hubert Index (HI) and Mirkin Index (MI) (Hubert and Arbie 1985). It is evident from Table 1 that the proposed algorithm provides good performance and thus it produces better consensus from multiple crowdsourced clustering solutions.

Table 1: Performance metric values for balance scale dataset.

Algorithm	Adjusted Rand	Rand	MI	HI
CSPA	0.1579	0.5941	0.4059	0.1883
HGPA	0.1482	0.5216	0.4784	0.0431
MCLA	0.0118	0.5985	0.4015	0.1971
BCE	0.0830	0.5616	0.4384	0.1232
Proposed	0.1767	0.6071	0.3929	0.2141

Conclusions

In this paper, a crowdsourcing model for grouping a set of tricky images is introduced and a Markov chain based ensemble method has been proposed. It can be adopted as an effective tool to achieve a good consensus from multiple diverse clustering solutions. Furthermore, it can also be extended to work with input crowdsourced clustering having variable number of clusters instead of fixed number of clusters considering other features (e.g., confidence and bias) of crowd workers.

Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable comments that greatly helped to improve the quality of the paper. All the authors would like to thank the crowd contributors involved in this work.

References

- [Ayad and Kamel 2008] Ayad, H. G., and Kamel, M. S. 2008. Cumulative voting consensus method for partitions with variable number of clusters. *IEEE transactions on pattern analysis and machine intelligence* 30(1):160–173.
- [Berkhin 2002] Berkhin, P. 2002. A survey of clustering data mining techniques. 25–71.
- [Brabham 2013] Brabham, D. C. 2013. Detecting stable clusters using principal component analysis. *Methods Mol. Biol.* 224(10).
- [Chatterjee and Mukhopadhyay 2013] Chatterjee, S., and Mukhopadhyay, A. 2013. Clustering ensemble: A multiobjective genetic algorithm based approach. In *Proceedings of International Conference on Computational Intelligence: Modeling, Techniques and Applications (CIMTA)*, 443–449. Procedia Technology, Elsevier.
- [Davies and Bouldin 1979] Davies, D. L., and Bouldin, D. 1979. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1:224–227.
- [Dunn 1974] Dunn, J. 1974. Well separated clusters and optimal fuzzy partitions. *J Cyberns* 4:95–104.
- [Hovy et al. 2013] Hovy, D.; Kirkpatrick, T. B.; Vaswani, A.; and Hovy, E. 2013. Learning whom to trust with mace. In *Proceedings of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL HLT)*, 1120–1130.
- [Hubert and Arbie 1985] Hubert, L., and Arbie, P. 1985. Comparing partitions. *Journal of Classification* 2(1).
- [Jain, Murthy, and Flynn 1999] Jain, A. K.; Murthy, M. N.; and Flynn, P. J. 1999. Data clustering: A review. *ACM Computing Surveys* 31:264–323.
- [Lease 2011] Lease, M. 2011. On quality control and machine learning in crowdsourcing. In *Proceedings of 3rd Human Computation Workshop (HCOMP) at AAAI*, 97–102.
- [Mukhopadhyay, Maulik, and Bandyopadhyay. 2015] Mukhopadhyay, A.; Maulik, U.; and Bandyopadhyay, S. 2015. A survey of multiobjective evolutionary clustering. *ACM Computing Surveys* 47(4):61:1–61:46.
- [Rand 1971] Rand, W. M. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* 66(336):846–850.
- [Strehl and Ghosh 2002] Strehl, A., and Ghosh, J. 2002. Cluster ensembles - a knowledge reuse framework for combining partitionings. *Journal of Machine Learning Research* 3:583–617.
- [Wang, Shan, and Banerjee 2009] Wang, H.; Shan, H.; and Banerjee, A. 2009. Bayesian cluster ensembles. In *Proceedings of the Ninth SIAM International Conference on Data Mining*, 211–222.